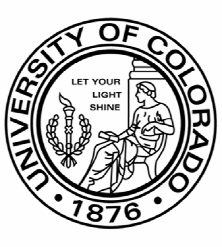
The Blue Gene/L System at NCAR
A Catalyst for Petascale Science

Henry Tufo

Scientist III and Computer Science Section Head
Computational and Information Systems Laboratory
National Center for Atmospheric Research

Associate Professor and Director, Computational Science Center
Department of Computer Science
University of Colorado



NCAR

Motivation

- There is a broad spectrum applications capable of producing groundbreaking science with the 1-2 order magnitude increase in computational power expected in the coming decade.
- The Computer Science Section is tasked with exploring new technologies, creating a cyberinfrastructure that integrates these technologies and increases scientific productivity, and developing the next generation of scientific models.
- This talk focuses on our evaluation of IBM Blue Gene (BG) technology, integration of BG/L into the TeraGrid, and the development of a version of the Community Climate System Model (CCSM) capable of exploiting the petascale class BG systems coming on line in the next 2-5 years.

Take Home Points

- On track to deliver the next version of CCSM that is capable of exploiting petascale systems with O(10K) to O(100K) cores.
- Shown that Blue Gene/L (BG/L) can be integrated into NSF's TeraGrid initiative (without too much difficulty).
- Added to BG/L the ability to efficiently accommodate single processor jobs with little modification to the software stack.

Outline

- NCAR's TeraGrid Node
- High Throughput Computing (HTC) for BG/L
- Community Climate System Model (CCSM)
  - POP
  - CICE
  - CLM/MCT
  - HOMME
  - Petascale CCSM
- Acknowledgements / Conclusions / Further Information

NCAR's TeraGrid Node

NCAR TeraGrid Resources

- **NCAR** is currently connected to both the TeraGrid and National Lambda Rail (NLR) networks that support direct 10 Gb/s connectivity to several internal systems as well as a dedicated 10 Gb/s link to the University of Colorado. Within NCAR's 10 Gb/s network, systems are connected with 10 Gb/s Ethernet as well as 802.3ad link aggregates of several 1 Gb/s Ethernet links. Current resources:

- **Tengee:** Our local 10 Gb/s Ethernet test host. With this host we have been able to sustain 9.8 Gb/s of TCP throughput, and we have used it to tune and benchmark our other systems and our 10 Gb/s links to CU and the TeraGrid.

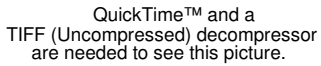
- **Twister:** A Sun server that provides a platform for the VAPOR data visualization project. Twister reads data from GPFS-WAN over the TeraGrid, generates visualizations, and sends them to remote displays at CU and other sites. This prototype project may be developed into a future TeraGrid science gateway.

- **Datagrid:** A Sun V890 server used to stage data for the Community Data Portal (CDP) and Earth System Grid (ESG) projects. Datagrid collects data from archival storage at NERSC and ORNL as well as from the NCAR Mass Storage System and makes it available to researchers worldwide via a TeraGrid science gateway.

- **Maelstrom:** A parallel storage cluster currently supporting Lustre, PVFS2, and experimental file systems. The Maelstrom cluster is composed of 14 individual servers and storage in a high-availability configuration, with an aggregate raw capacity of 120 TB. Maelstrom is used for wide-area Lustre and other TeraGrid storage projects.

- **Frost:** A single-rack IBM BG/L system with associated storage and login nodes. The BG/L rack has 2,048 700-MHz PowerPC 440 CPUs and 512 GB of RAM, with a peak performance of 5.7 teraflops. The storage and login nodes are four IBM OpenPower 720 systems, each of which offers four POWER5 1.65-GHz CPUs and 8GB of RAM. Together, they provide 6 TB of local scratch space. One quarter of Frost will be available to TeraGrid users at the end of July 2007.

NCAR TeraGrid Integration



Frost Integration as a TeraGrid Resource

- Production TeraGrid resource at 25% level as of 8/1/2007.
- Several tasks required for Frost's integration:
  - Deployment of resources on TeraGrid network
  - Hardening security for BGL front-end nodes
  - Wide-area file system integration and analysis (GPFS-WAN)
  - Wide-area network validation and testing
  - Common TeraGrid Software Stack (CTSS) deployment
  - NCAR accounting and TGCDB integration
  - Integrating local resource manager with CTSS

Frost Integration Experiences

- Reconfiguration of Frost to properly integrate with the TeraGrid was a challenge:
  - Unfamiliar and complex software stack
  - Shift in security, operating, and usage policies
- Phased deployment eased integration:
  - Started with a single BG/L front-end node
  - Gradually integrated other nodes once policies and reconfiguration procedures refined from test deployment
  - Followed with deployment of CTSS and NCAR accounting integration
  - Concluded with friendly user testing

HTC on BG/L

HTC Motivation

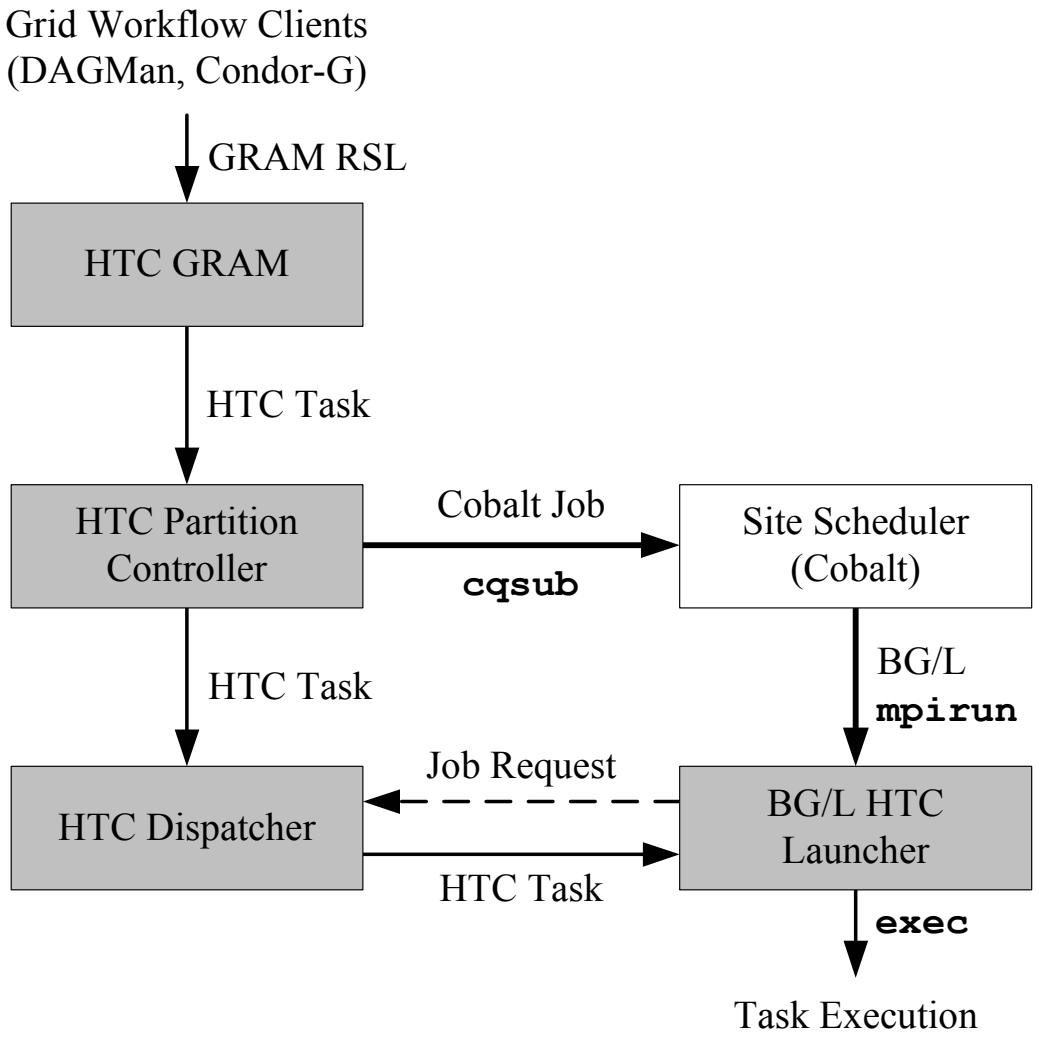
- Emerging high-performance computing architectures, such as BG/L, are designed for massively parallel applications.
- NCAR's role as a resource provider for the Earth Science community has embraced these architectures:
  - Large quantity of computing elements
  - Modest space and power constraints
  - Reasonable cost for performance
- NCAR's role as a TeraGrid Resource Provider will expose these resources to different application and workflow requirements:
  - Large user group with a variety of serial and parallel applications.
  - Require tools to provision and manage NCAR resources that can accommodate these workloads.
  - **These resources do not adequately support workflows composed of many serial tasks.**

HTC on BG/L

Solution: Develop infrastructure to support the execution and management of serial and parallel Grid-enabled workflows for Blue Gene/L

- Leverage existing technology and infrastructure:
  - BG/L High Performance Computing mode (std.)
  - BG/L High Throughput Computing mode
    - Permits the execution of non-MPI based tasks on the BGL
    - Uses a simple client / server architecture
  - Local resource managers (Cobalt)
  - Grid computing tools (Globus Toolkit, Condor-G, CTSS, etc.)
- Simultaneously support massively parallel and serial workflows
- Minimize system reconfiguration
- (Related work integrates HTC mode to support Condor on BG/L. Requires running Condor daemons on BGL front-end and I/O nodes.)

Design and Implementation

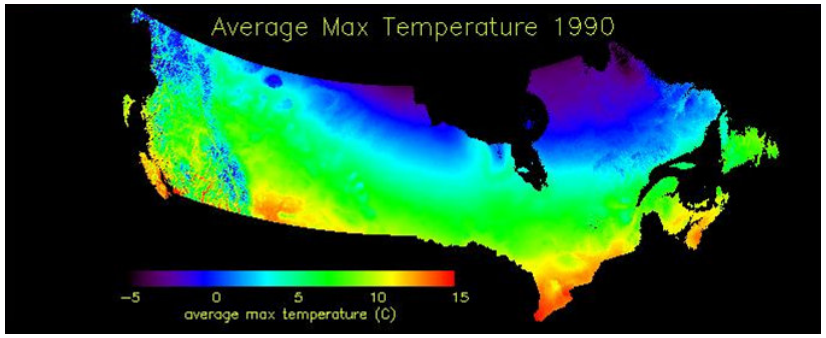


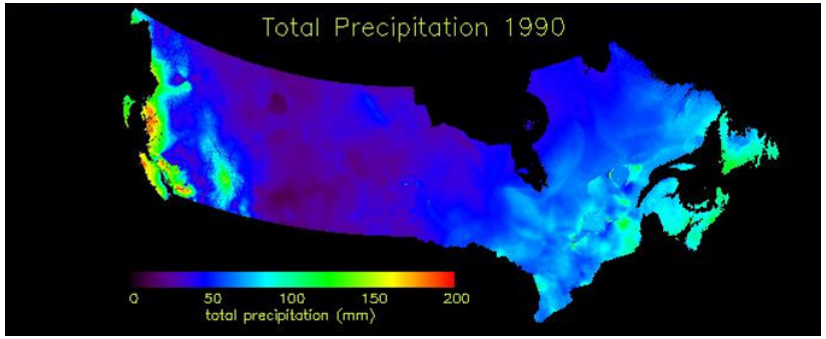
Evaluation

- Evaluated HTC using several synthetic workloads:
  - Simulated workloads of many serial tasks
  - Evaluated up to 512 simultaneous job submissions
  - Varied executable size
- Overhead introduced by HTC is minimal:
  - HTC submission interface scales well
  - Job throughput from a cold start is dependent upon the number of partitions required to execute the tasks
  - Possible contention on the BG/L IO node as the number of simultaneous serial tasks increases

Using HTC To Execute Grid Workflows - Grid-BGC

- Grid-BGC is a grid enabled tool for simulating the carbon cycle:
  - Science Portal
  - Service Oriented Architecture (data, workflow, job execution, etc.).
- Ported Daymet to BG/L:
  - Regional model that predicts daily and annual temperature, precipitation, and solar radiation amounts
  - First step of the Grid-BGC workflow
- Daymet workflow profile:
  - Composed of 21 executables
  - Individual task profiles vary
  - Embarrassingly parallel



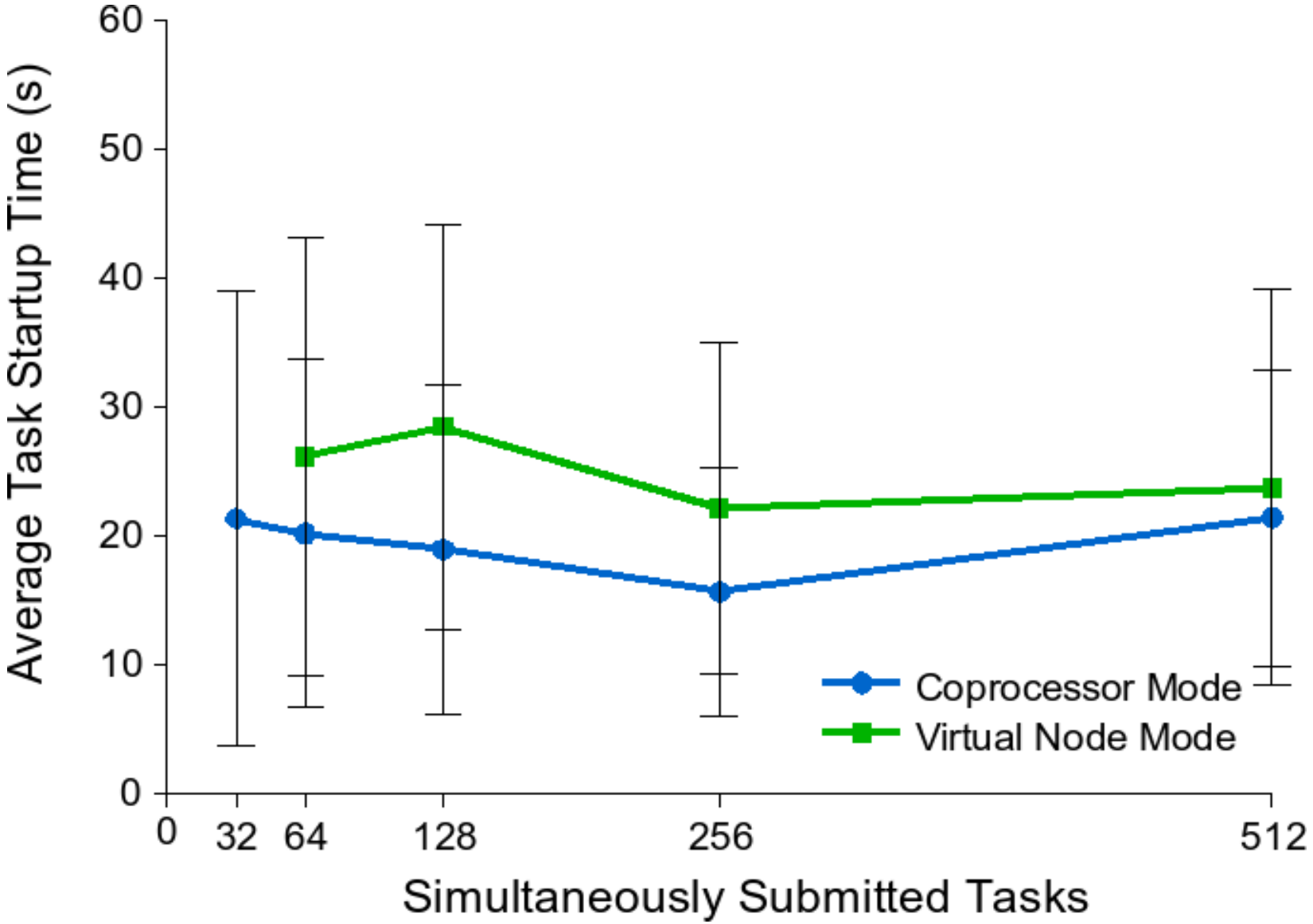


Executing Grid Workflows

- Daymet workflows and BG/L HTC:
  - Derived from workflows submitted to Grid-BGC
  - Serial workflows using Globus GRAM and Condor-G clients
  - Parallel workflows using Condor-G / DAGMan clients

- Successfully executed Daymet simulation of the continental United States for 2004:
  - 263 workflows, 5523 workflow nodes / execution tasks
  - Executed in approximately 10 hours on 32 compute nodes

HTC Performance

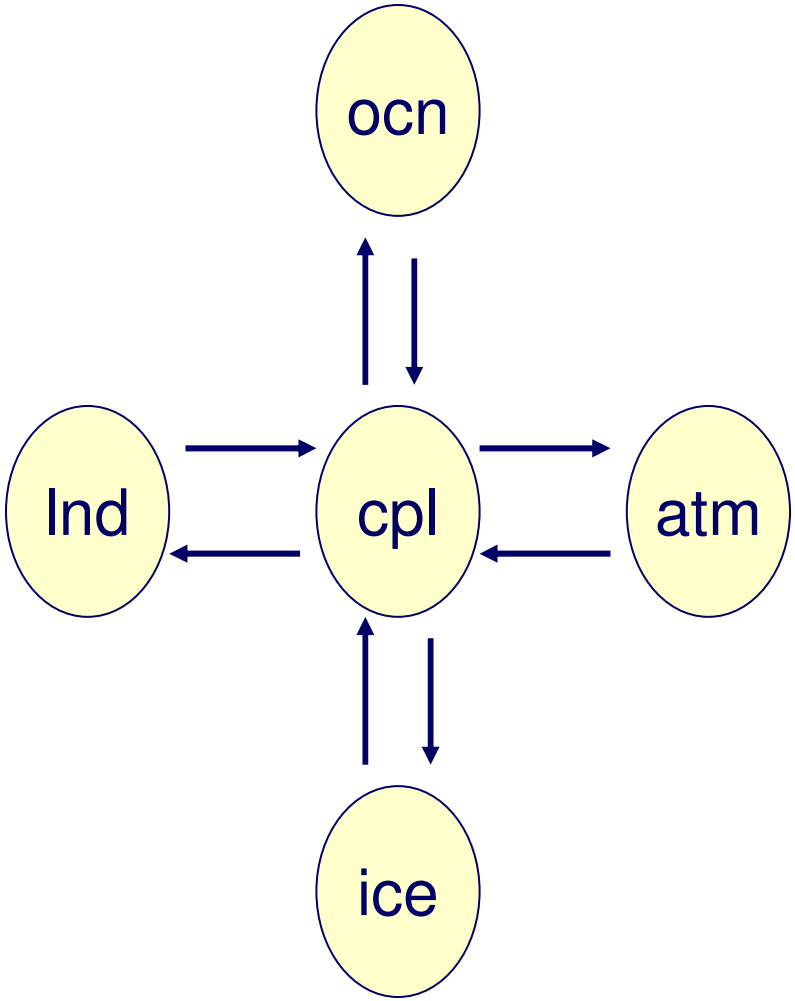


Shameless Plug

- J. Cope, M. Oberg, H.M. Tufo, T. Voran, M. Woitaszek, “High Throughput Grid Computing with an IBM Blue Gene/L”, Proceedings of IEEE Cluster 2007.
- Come see the full talk at Cluster 2007 in Austin, TX.

CCSM

CCSM Design and Details



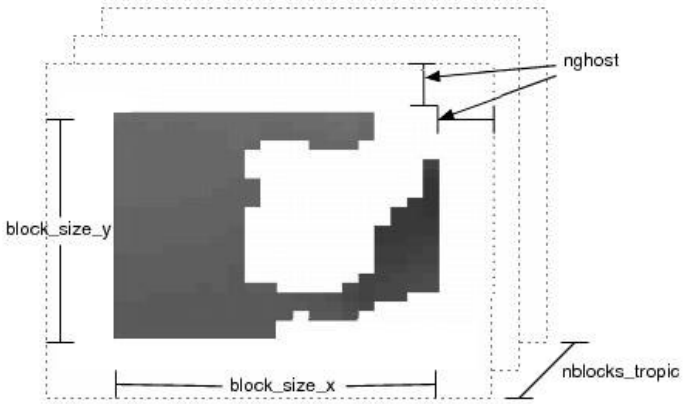
- Hub and spoke design
- Multiple executables (60K-200K lines each)
  - Ocean (ocn): POP
  - Atmosphere (atm): CAM/HOMME
  - Sea Ice (ice): CICE
  - Land (lnd): CLM
  - Coupler (cpl): MCT
  - (Single executable in beta.)
- Need 5 simulated years/day implies that we must run at "low" resolution.
- Typical configuration run on O(200) processors. Key question is whether we can scale up the individual components without adding work.
- Target Petascale Configuration:
  - CAM - 30 km, L26
  - POP, Sea Ice, and Land - 0.1°

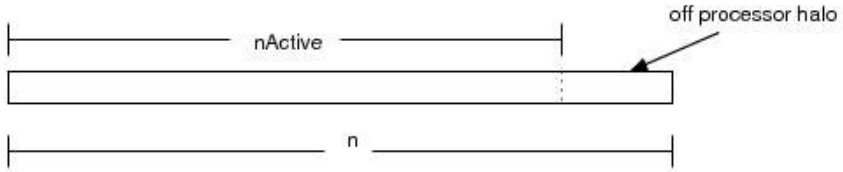
POP

POP (Parallel Ocean Program)

- Developed at LANL
- Two components:
  - Baroclinic: Finite difference
  - Barotropic: Solve surface pressure (2D) with PCG (diagonal preconditioning)
- John Dennis modified base POP 2.0 base code to:
  - Reduce execution time/ improve scalability,
  - Minor changes (to ~9 files) yields 2x performance improvements!
    - Redesign of the barotropic solver using 1D data structures.
    - Aggregated 3-D boundary exchange.
    - Improve partitioning/load-balancing using space-filling curves.

New 1D Data Structure





2D data structure

- Advantages
  - Regular stride-1 access
  - Compact form of stencil operator
- Disadvantages
  - Includes land points
  - Problem specific data structure

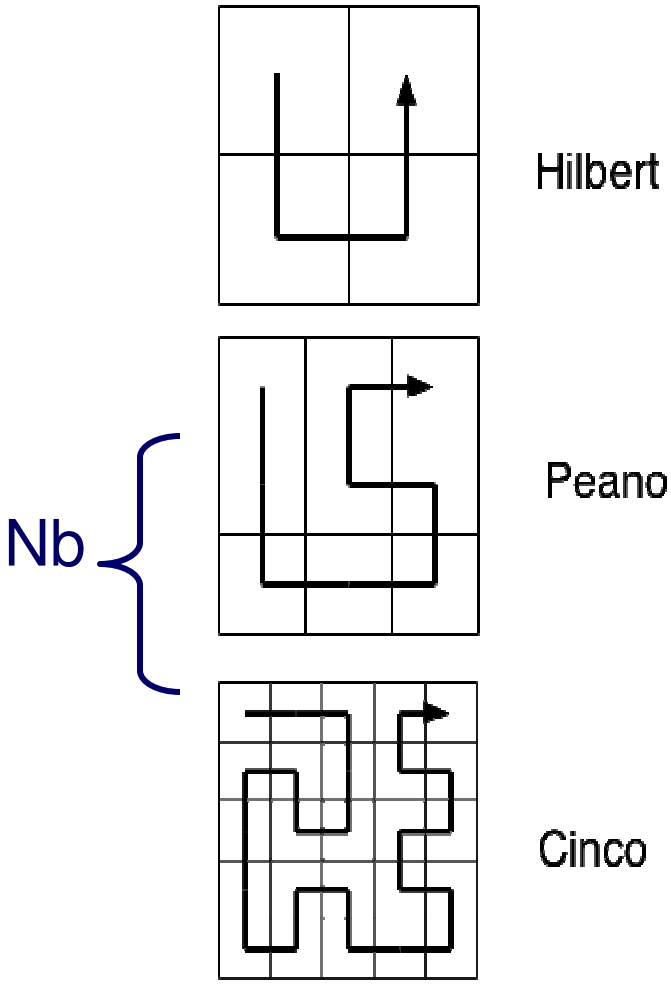
1D data structure

- Advantages
  - Removes land points
  - General data structure
- Disadvantages
  - Indirect addressing
  - Larger stencil operator

Aggregated Boundary Exchange

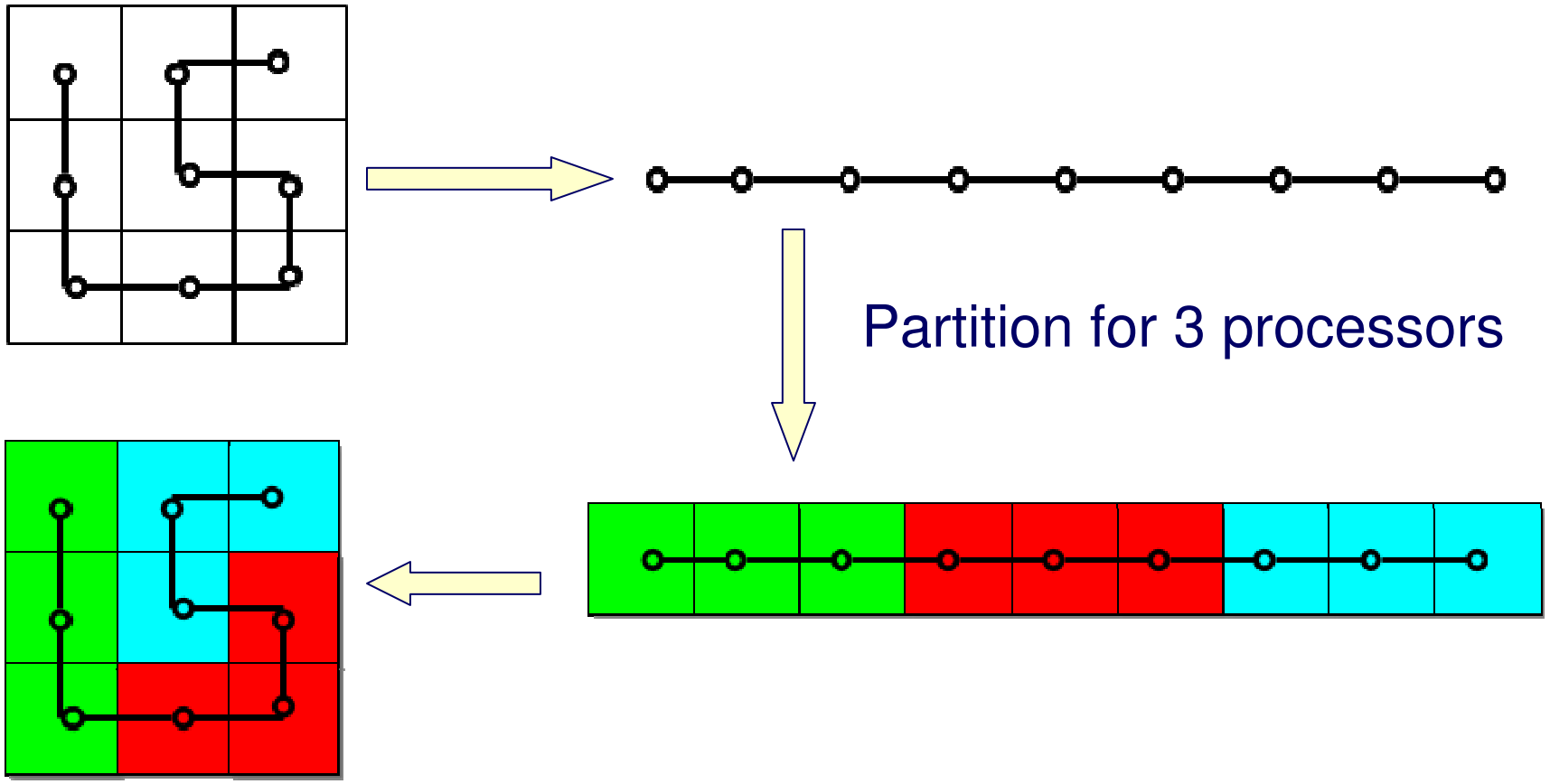
- Originally, POP applied 2D boundary exchange to 3D variables. 40 layers implies 40 boundary exchanges.
  - Increases latency sensitivity!
- 3D-update can consume up to 33% of total execution time.
- Developed a specialized 3D boundary exchange:
  - Reduces message count
  - Increases message sizes
  - Reduces dependence on machine latency
- Imported implementation from CICE 4.0 boundary exchange.

Space-Filling Curves

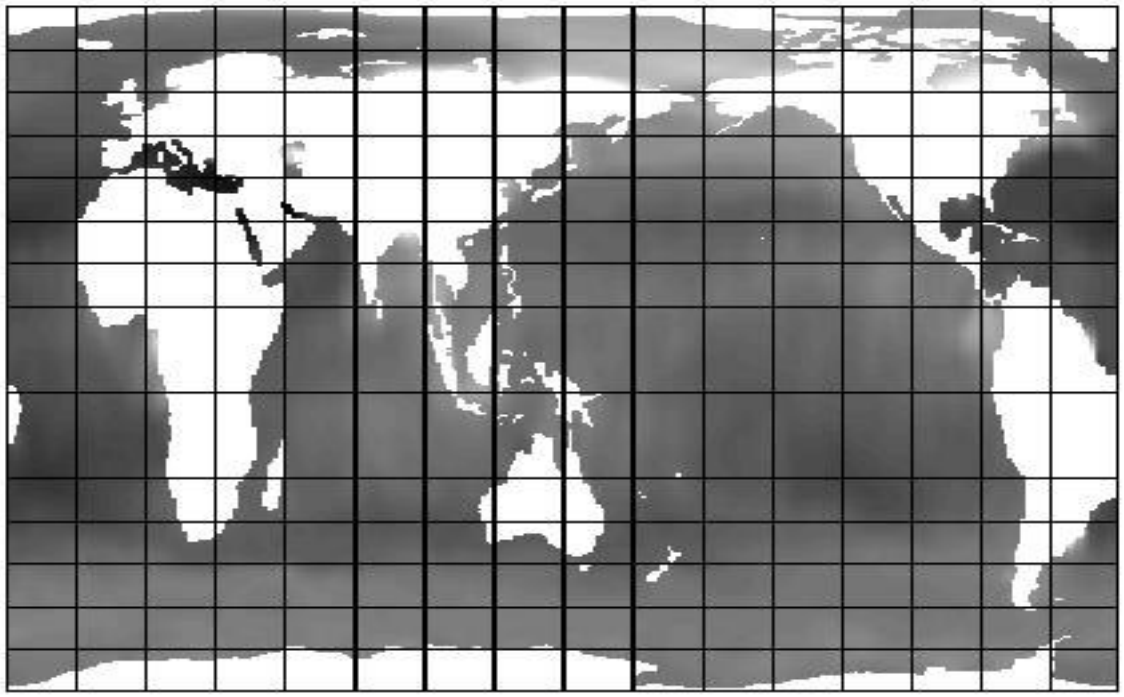


- Use SFC map 2D objects into 1D.
- Variety of sizes
  - Hilbert ($Nb=2^n$)
  - Peano ($Nb=3^m$)
  - Cinco ($Nb=5^p$) [New]
  - Hilbert-Peano ($Nb=2^n3^m$)
  - Hilbert-Peano-Cinco ($Nb=2^n3^m5^p$) [New]
- Partitioning 1D array

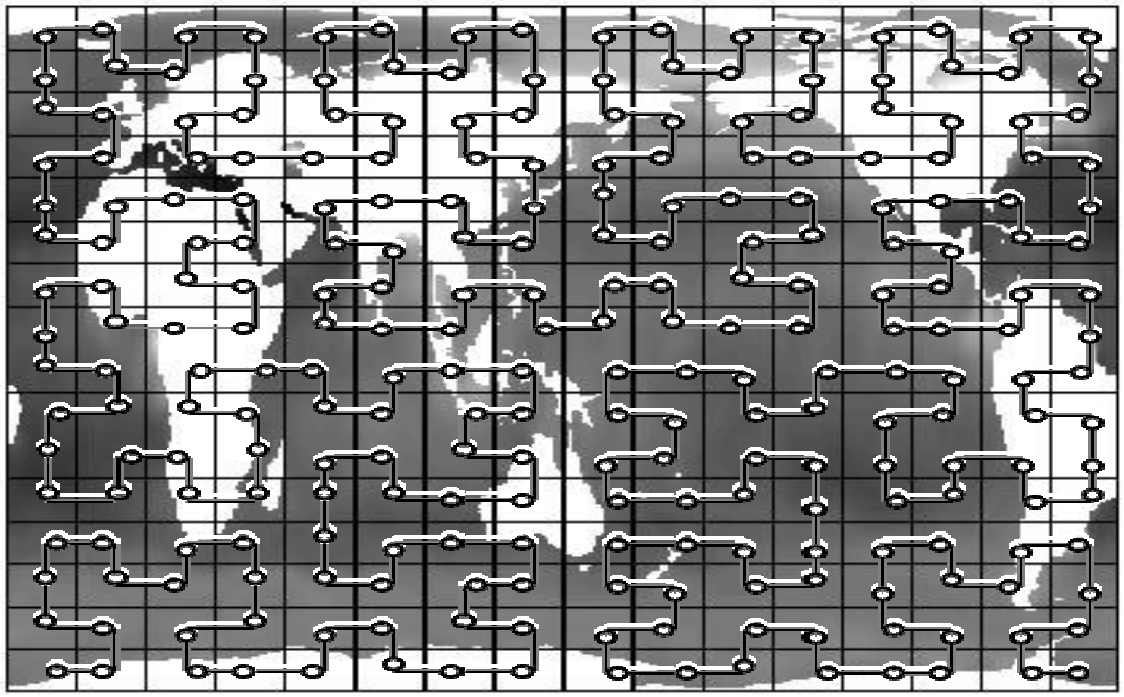
Partitioning With SFC



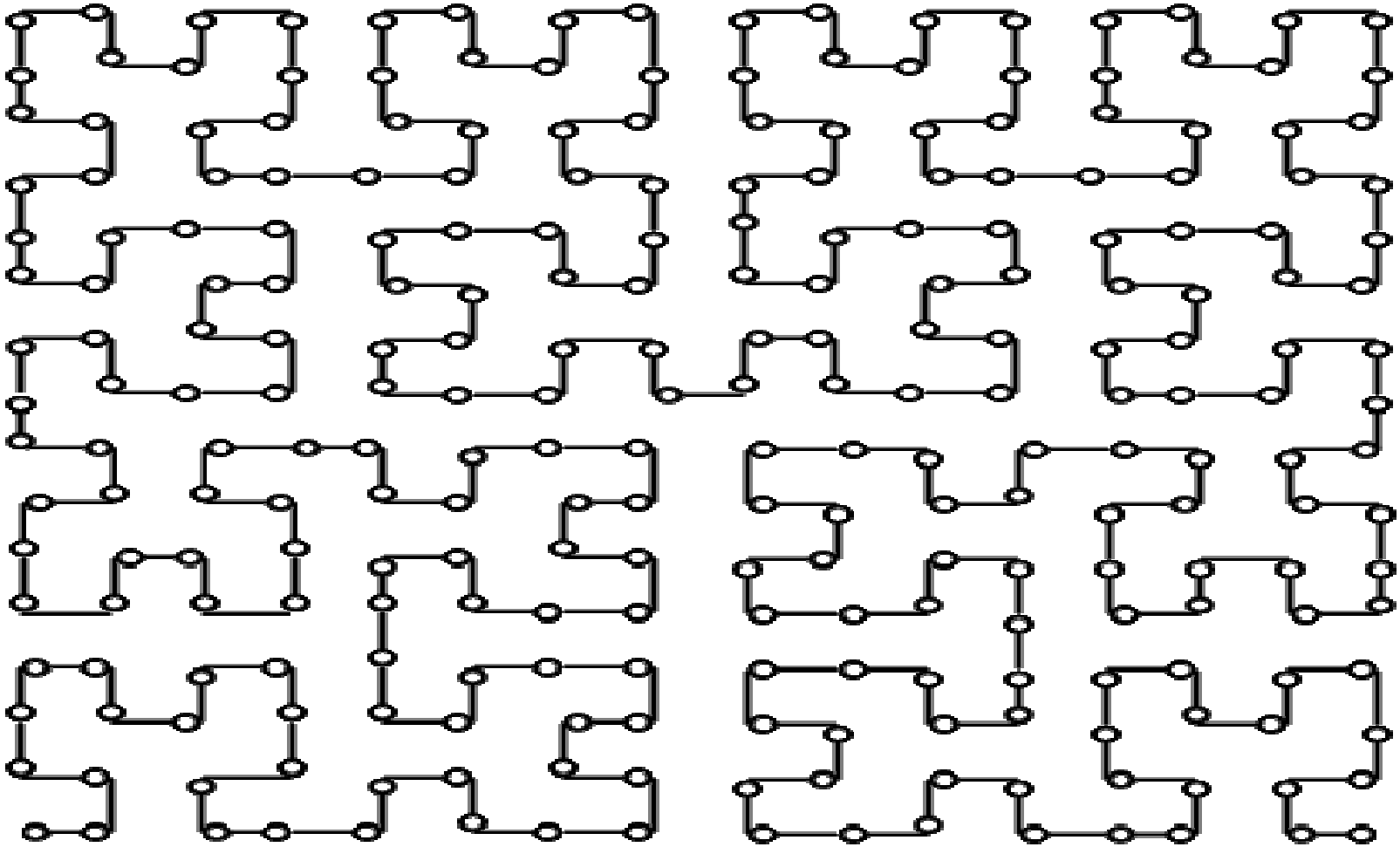
POP Using 20x24 blocks (gx1v3)



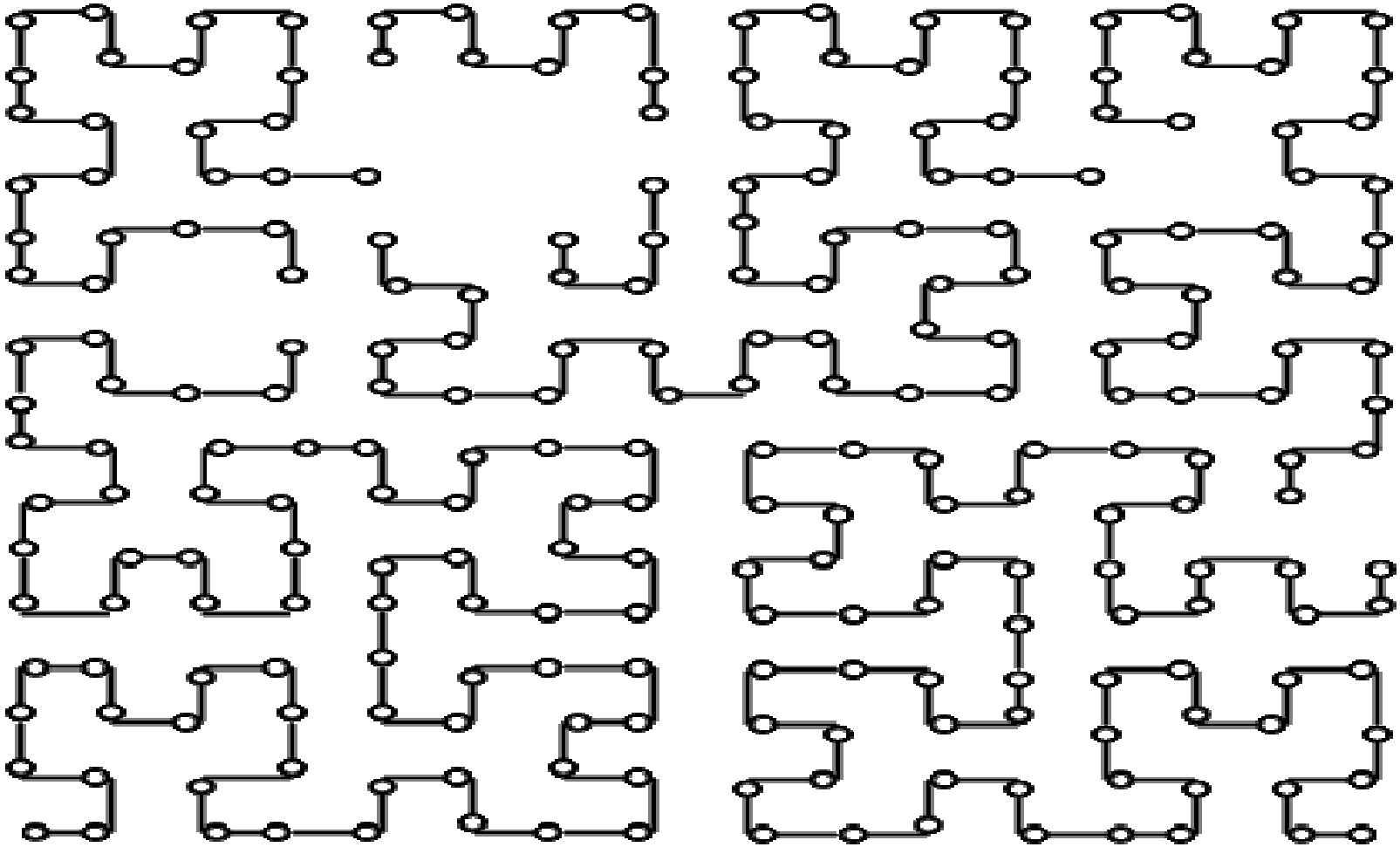
POP (gx1v3) + Space-Filling Curve



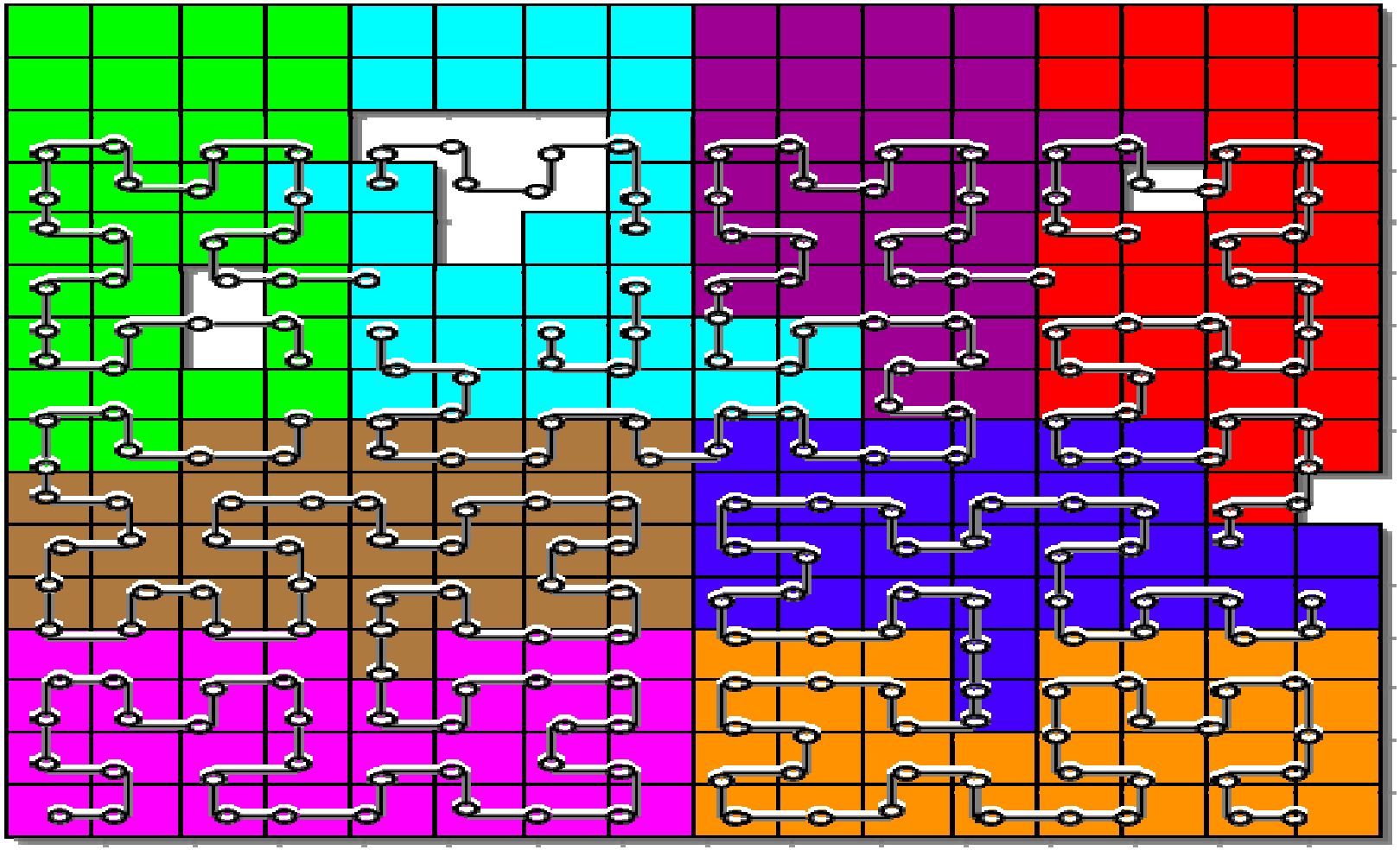
Space-Filling Curve (Hilbert Nb=2^4)



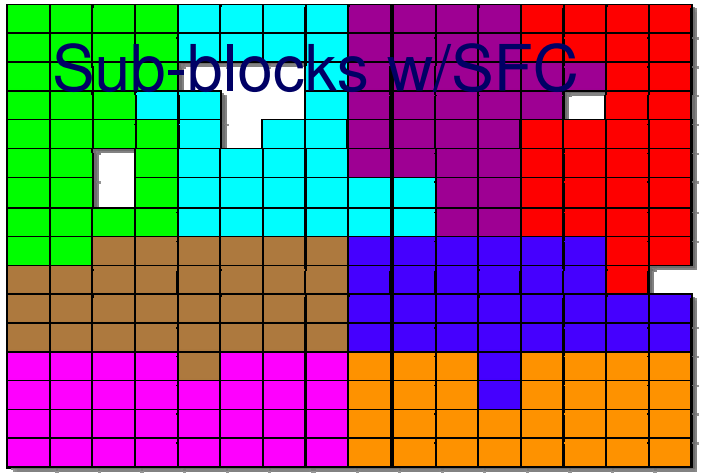
Remove Land Blocks

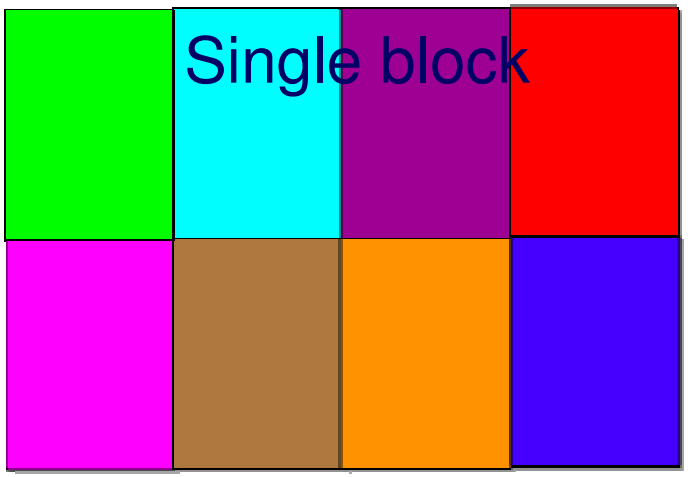


Space-Filling Curve Partition for 8 Processors



Decomposition on 8 Processors



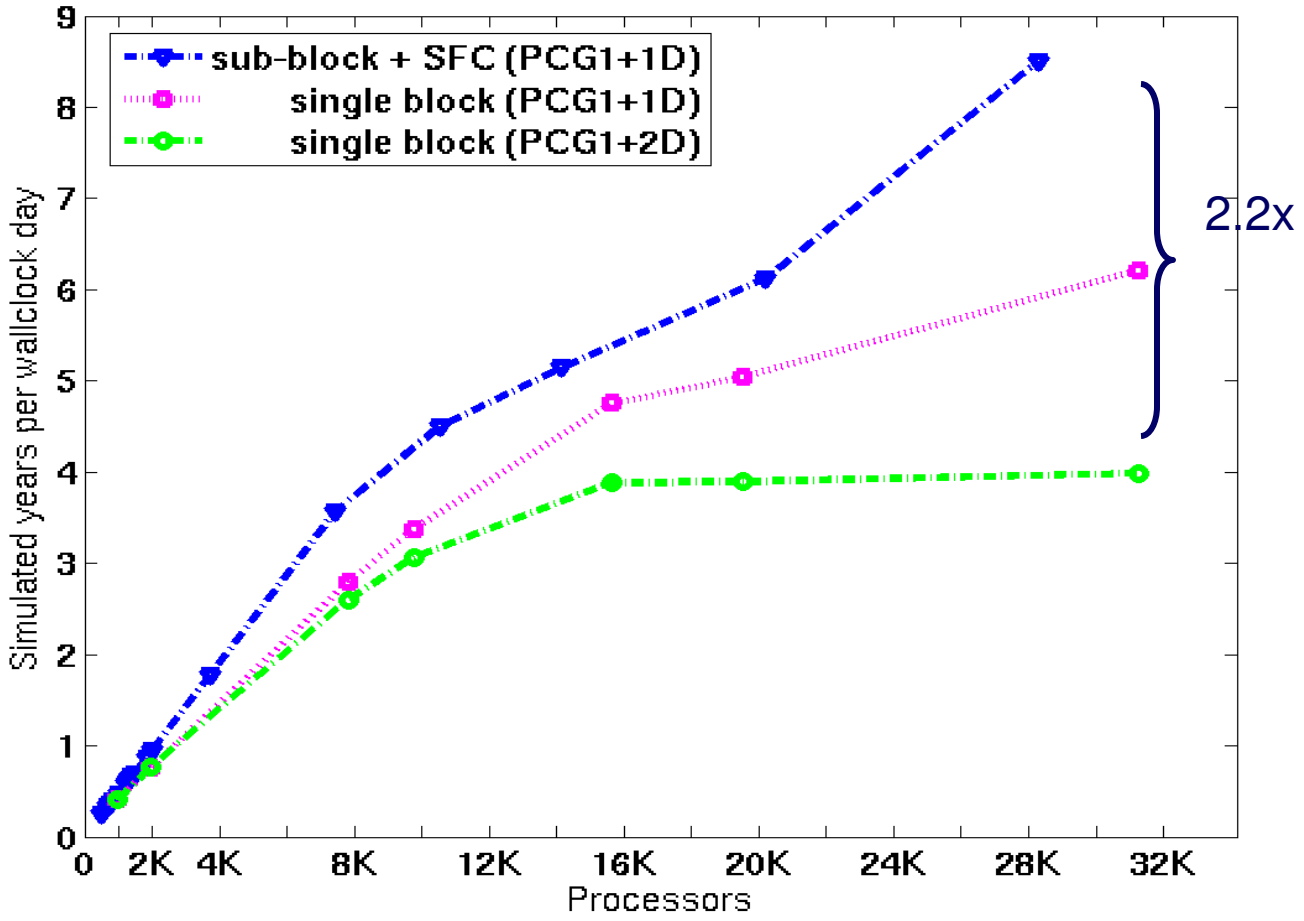


- Sub-block w/SFC partitioning
  - Improved load balanced [eliminate land blocks]
  - May increase communication volume
- Single block
  - Includes all land points
  - Idle processors are possible

Test Case - 0.1° POP

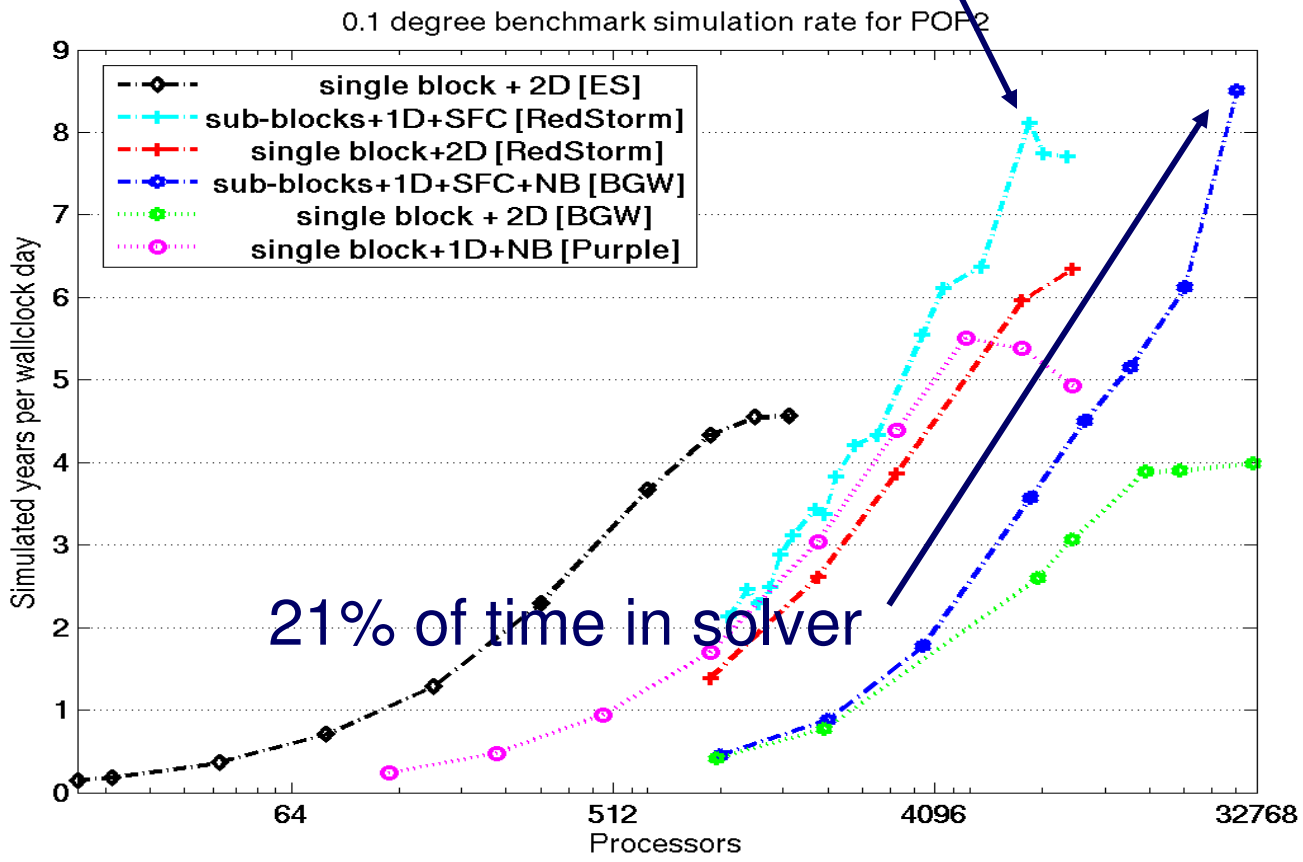
- Global eddy-resolving
- Computational grid:
  - 3600 x 2400 x 40
- Land creates problems:
  - Load imbalances
  - Scalability
- Alternative partitioning algorithm:
  - Space-filling curves
- Evaluate using Benchmark:
  - 1 day/ Internal grid / 7 minute time step

POP - 0.1° Benchmark on BG/L



POP - 0.1° Benchmark

50% of time in solver



21% of time in solver

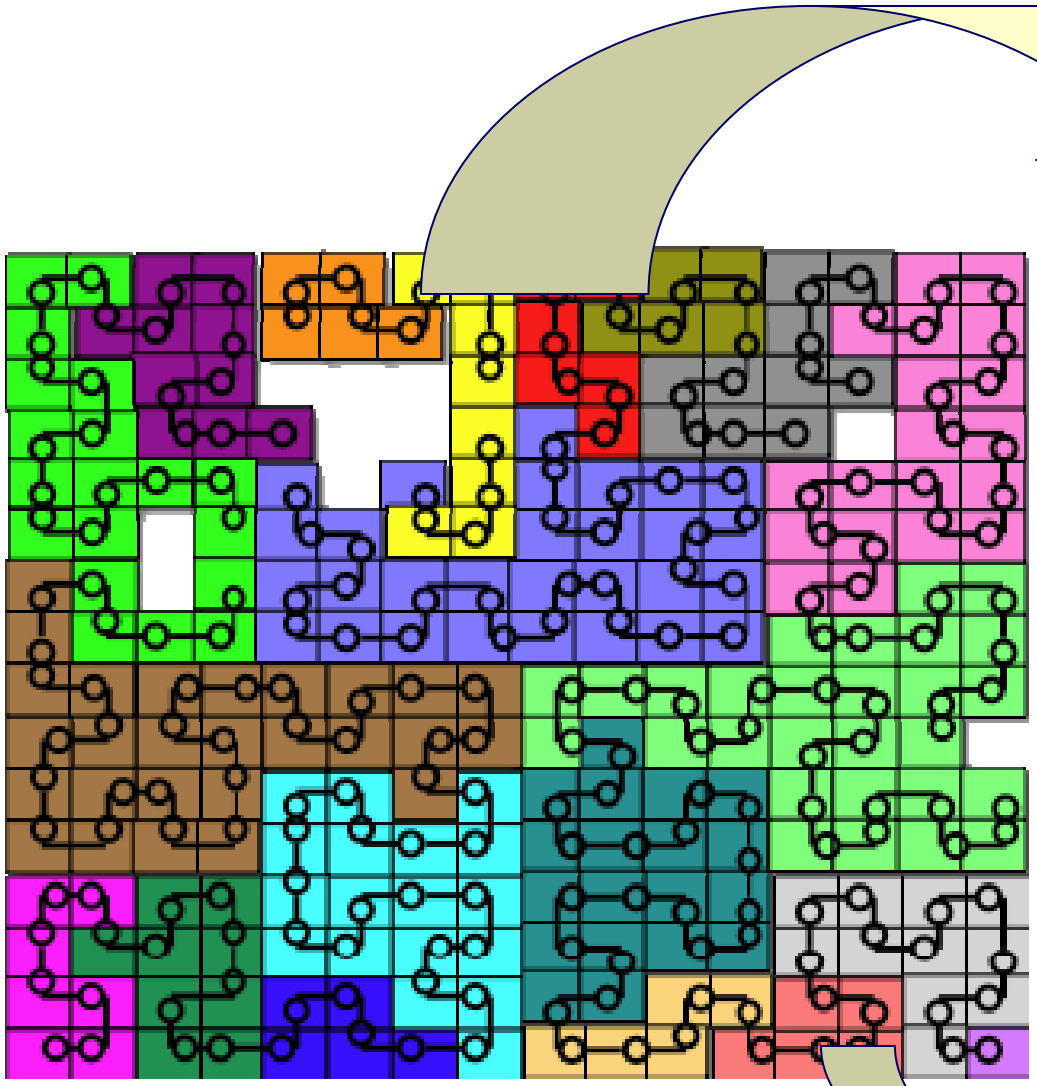
Courtesy of Y. Yoshida, M. Taylor, P. Worley

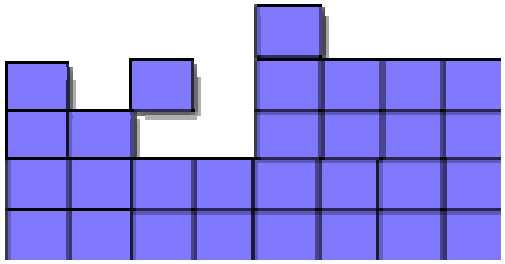
CICE

CICE4 - 0.1°

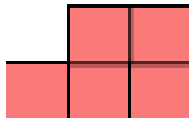
- Developed at LANL
- Finite difference
- Models sea ice
- Shares grid and infrastructure with POP
  - Reuse techniques from POP work
- Computational grid: [3600 x 2400 x 20]
- Computational load-imbalance creates problems:
  - ~15% of grid has sea-ice
  - Try weighted space-filling curves
- Evaluate using Benchmark:
  - 1 day/ Initial run / 30 minute time step/ no Forcing
  - 10K Cray XT3 processors
  - 40K Blue Gene/L processors

1° CICE4 on 20 processors



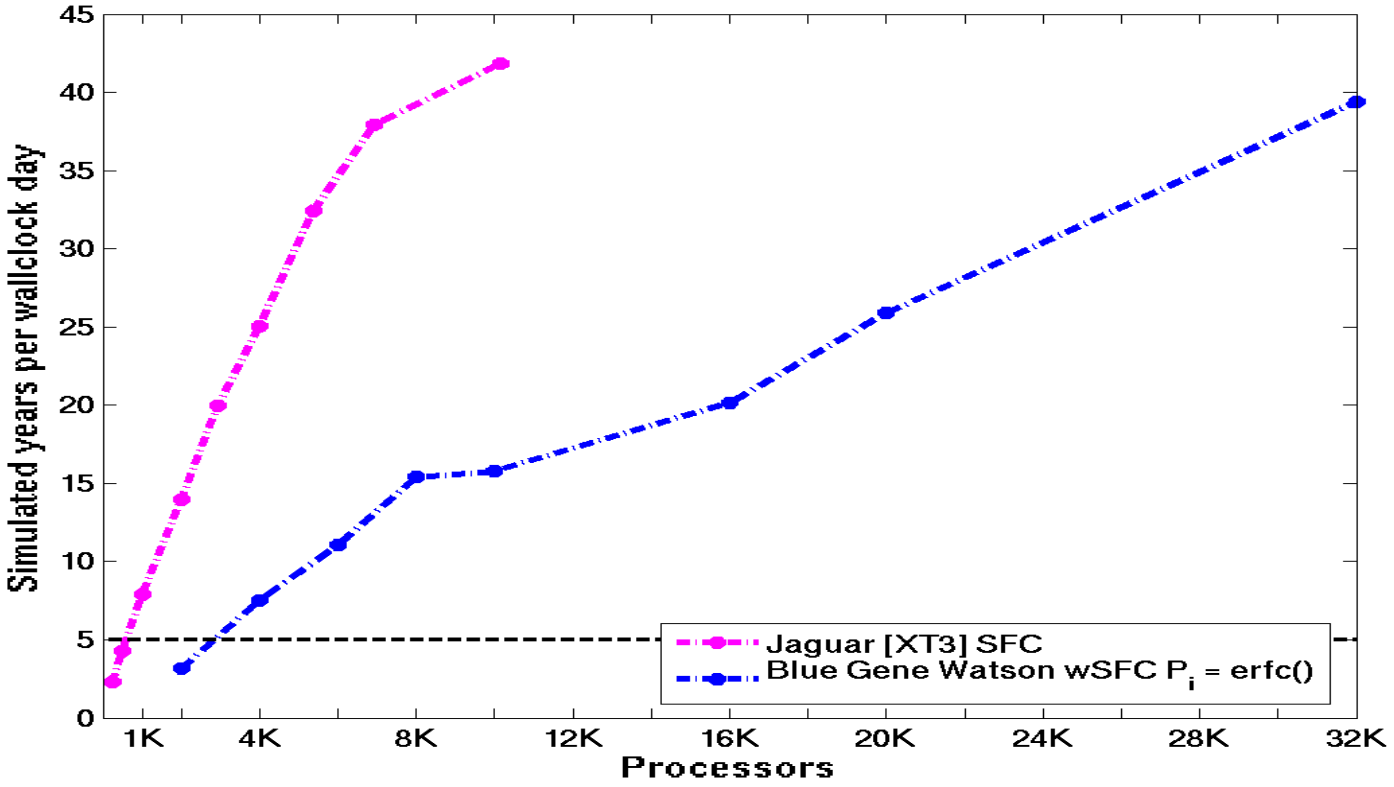


Large domains @ low latitudes



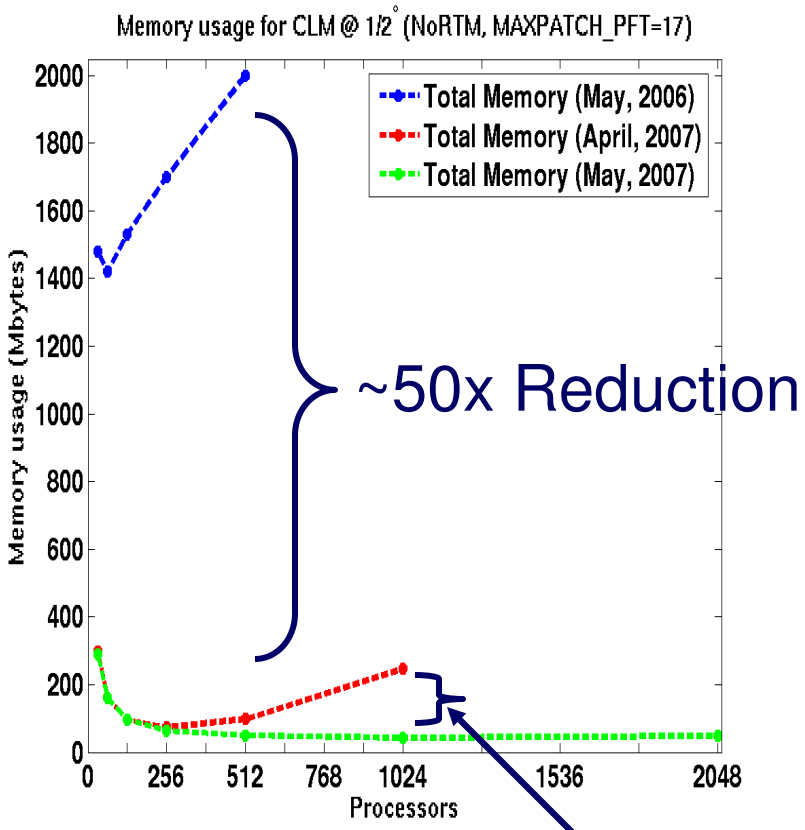
Small domains @ high latitudes

CICE4 @ 0.1°


Simulated years per wallclock day
45
40
35
30
25
20
15
10
5
0
1K
4K
8K
12K
16K
20K
24K
28K
32K
Processors
Jaguar [XT3] SFC
Blue Gene Watson wSFC P_i = erfc()

CLM/MCT

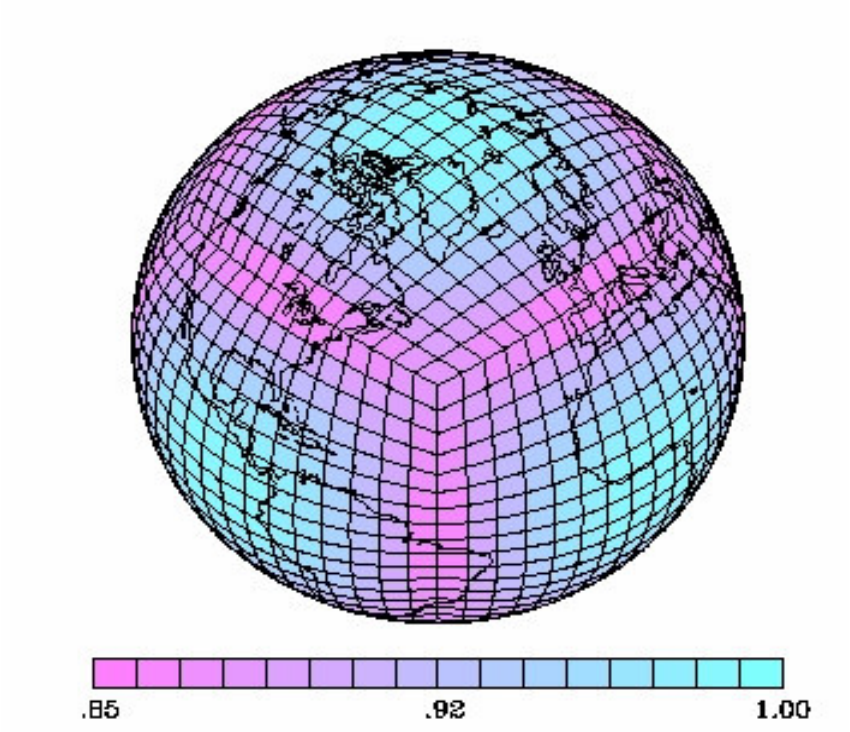
Status of CLM/MCT



- Work of T. Craig
- Elimination of global memory
  - Reworking of decomposition algorithms
- Resources for 1/6° global run
  - May 2006: 512 processors, ~18 Gbytes per proc
  - May 2007: 512 processors, ~495 Mbytes per proc
- Future Work
  - Investigation scalability at 1/6° & 1/10°
  - Addition of PIO

HOMME

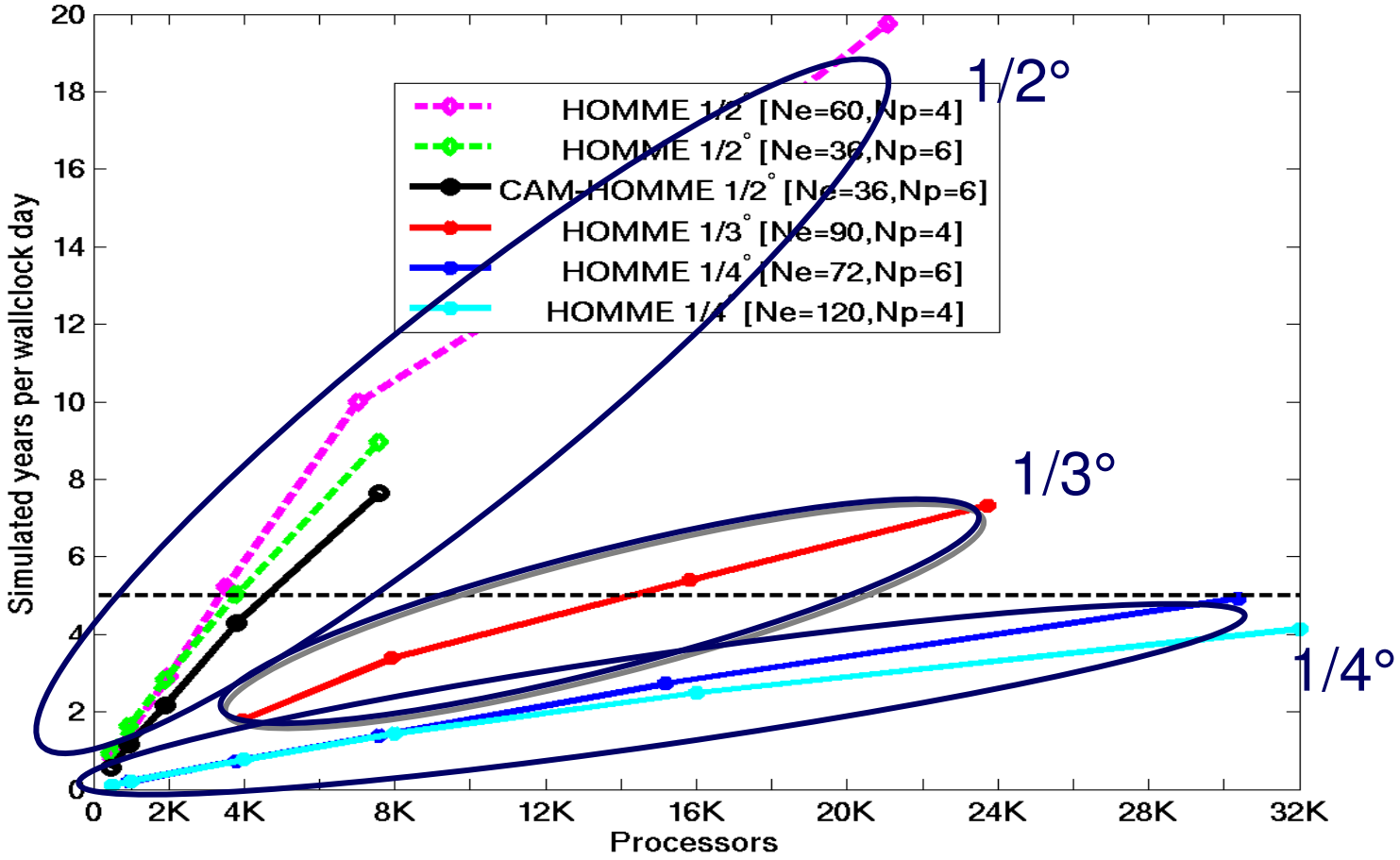
High-Order Method Modeling Environment (HOMME)



Ne=16 Degree of non-uniformity

- Cube-Sphere Mesh
- Supports Spectral Elements (SE) and Discontinuous Galerkin (DG)
- HOMME Benchmark
  - 30km [13824 procs]
  - 10km [98K procs]
- Demonstrated Scalability
  - 32K on Blue Gene
  - 10K on XT3
- Alternative Dynamic Core in CAM
  - DOE SciDAC-CCPP, Tufo PI
  - Ram Nair HOMME DG Lead
  - Mark Taylor HOMME SE Lead

HOMME Simulation Rate - Held-Suarez



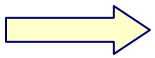
Petascale CCSM

Ultra-high resolution CCSM Simulation Blue Gene/P (BG/P)

- ~Petascale BG/P system (2008)
  - 160K processor system
  - 512Mbytes/proc
  - 4 processors per node
- Petascale Configuration:
  - CAM (30 km, L26)
  - POP @ 0.1°
  - Sea-Ice @ 0.1°
  - Land model @ 0.1°
- Duration: 30 year run
- Secondary storage: ~12 Tbytes per run
- Tertiary storage: ~100 Tbytes per run

Feasibility Study

	IBM Blue Gene L			Cray XT3		
	nprocs	Measured	status	nprocs	Measured	status
0.1° POP	32,768	27.9 sec	done	17,000	20.0 sec	done
0.1° CICE	6,000	21.5 sec	done	1000	27.7 sec	done
0.1° CLM	4,096		[3Q07]	1,000		[3Q07]
CAM (30 km, 26 levels)	65,536		[2Q08?]	28,000		[??]
MCT	8192		[4Q07]	2,000		[4Q07]
Total	116,688	40 sec	[??]	49,00	40 sec	[??]



~5.9 years/wallclock day

Additional Details

- J. M. Dennis and H. M. Tufo, "Scaling Climate Simulation Applications on IBM Blue Gene", IBM Journal of Research and Design on Applications for Massively Parallel Systems. To appear.

Folks and Funding

- NCAR/CU Boulder
  - Ram Nair, Scientist II
  - John Dennis, Scientist I
  - Matthew Woitaszek, Scientist I
  - Jim Edwards, Software Engineer IV (IBM)
  - Sean McCreary, Software Engineer III
  - Adam Boggs, Software Engineer II
  - Rory Kelly, Software Engineer II
  - Michael Oberg, Software Engineer I
  - Hae-Won Choi, Postdoc
  - Jack Chen, Postdoc
  - Jason Cope, Graduate Student
  - Mike Levy, Graduate Student
  - Theron Voran, Graduate Student
  - Bobby House, Undergraduate Student
  - Eric Strohler, Undergraduate Student
  - Brandon Werdel, Undergraduate Student
  - Evan Gates, Undergraduate Student
- Collaborators:
  - Mark Taylor(Sandia), Joe Tribbia (NCAR), Dave Williamson (NCAR), Warren Washington (NCAR), Lawrence Buja (NCAR), John Drake (ORNL), Pat Worley (ORNL), Amik St-Cyr (NCAR), Steve Thomas (NCAR), Rich Loft (NCAR), Jan Mandel (CU Denver), Phil Rasch(NCAR)
- Funding from IBM, NASA, NSF, and DOE

Conclusions

- On track to deliver the next version of CCSM that is capable of exploiting petascale systems with O(10K) to O(100K) cores.
- Shown that Blue Gene/L (BG/L) can be integrated into NSF's TeraGrid initiative (without too much difficulty).
- Added to BG/L the ability to efficiently accommodate single processor jobs with little modification to the software stack.

Further Information

- tufo@ucar.edu, http://csc.cs.colorado.edu/~tufo/
- Computational Science Center: http://csc.cs.colorado.edu/
- Computer Science Section: http://www.cisl.ucar.edu/css/
- Blue Gene: http://csc.cs.colorado.edu/systems/frost/
- HOMME: http://www.homme.ucar.edu/
- DGAM: http://csc.cs.colorado.edu/SciDAC-CCPP/
- Grid-BGC: http://www.gridbgc.ucar.edu/